STAT 4410/8416 Homework 6

Islam MD Tahidul

Due on Dec 9, 2023

1. **Big data tools:** The Hadoop Distributed File System (HDFS) allows us to manipulate massive amount of data using scalable computing power. Please answer the questions below based on HDFS. You don’t have to show the results, just explain.

a. Explain what the following commands do.

```
hadoop fs -mkdir wordcount/input
hadoop fs -put myFile.txt myHdfs/test.dat
```

Answer: `hadoop fs -mkdir wordcount/input` creates a directory named input in the HDFS directory structure under the wordcount directory.
`hadoop fs -put myFile.txt myHdfs/test.dat` command copies the local file myFile.txt to the HDFS directory myHdfs with the name test.dat

b. Explain what the following `pig` commands will do.

```
dat = LOAD 'myHdfs/test.dat';
d = LIMIT dat 10;
DUMP d;
```

Answer: `dat = LOAD 'myHdfs/test.dat';` command loads data from the HDFS file test.dat located in the myHdfs directory into a Pig relation named dat.
`d = LIMIT dat 10;` command creates a new relation d selecting the first 10 records from the dat relation.
`DUMP d;` basically its a print command.

c. Write down two differences between `Pig` and `Hive`. Which code will run faster?
Answer:

- Pig operates on the client-side of a cluster whereas Hive operates on the server-side of a cluster.
- Pig does not have a dedicated metadata database whereas Hive makes use of the exact variation of dedicated SQL-DDL language by defining tables beforehand.

`Pig` code will run faster.

#Ref: https://www.projectpro.io/article/difference-between-pig-and-hive-the-two-key-components-of-hadoop-ecosystem/79# (https://www.projectpro.io/article/difference-between-pig-and-hive-the-two-key-components-of-hadoop-ecosystem/79#)

d. If a data manipulation process takes 10 days to complete, what can you do to finish it in one day?
Answe:I can utilize parallel computing.